Robust Ensemble Classifier Methods for Detection Problems with Unequal and Evolving Error Costs

Barry Y. Chen
(925) 423-9429
chen52@llnl.gov

Successful analysis in real-world detection applications often hinges upon the automatic collection of massive amounts of data over time. However, the pace of automatic data collection far exceeds our manual processing and analysis capabilities, making automated pattern detection in streaming data critical.

Machine learning classifiers capable of detecting patterns in datasets have been developed to address this need, but none can simultaneously address the many challenging characteristics of real-world detection problems. In particular, the costs associated with false alarms and missed detections are frequently unequal, extreme (demanding near-zero false alarm or miss rates), or changing over time. Moreover, the underlying data distribution modeled by the classifiers may also evolve over time, resulting in progressively degraded classification performance.

We are addressing these deficiencies via the development of new dynamic ensemble classifier algorithms that leverage diverse cost-sensitive base-classifiers.

Project Goals

The ultimate goal of this two-year effort focuses on the understanding and development of new ensemble learning algorithms that can effectively address the considerable challenges presented by detection problems of national significance. The developed methodologies will yield significantly improved performance at near-zero false alarm (or missed detection) rates and be able to adapt to changing costs and data distributions in a dynamic environment. Moreover, this research will lead to greater insight into the factors that interact to govern classification performance, including ensemble size, feature dimensionality, and data sampling.

Relevance to LLNL Mission

This research directly supports the Engineering Systems for Knowledge and Inference (ESKI) focus area and the Threat Prevention and Response Technologies theme in the LLNL Science and Technology Plan with an emphasis on knowledge discovery, advanced analytics, and architectures for national security. Our research explicitly addresses needs in the counterterrorism, nonproliferation, and national security missions for a broad range of customers, including the IC, DHS, DOE, DoD, and NNSA.

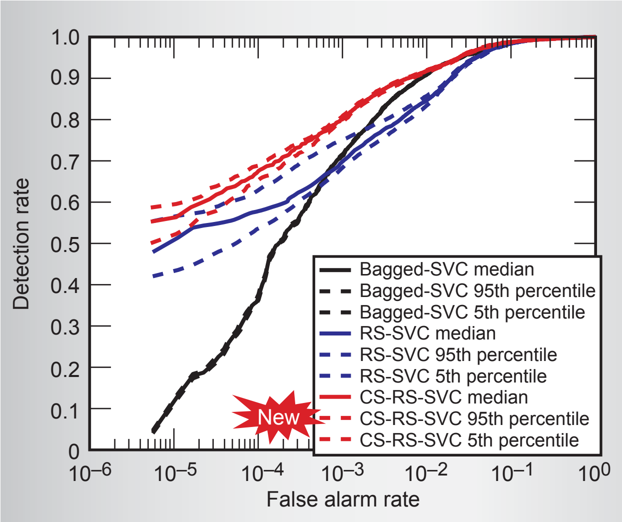

Figure 1. Median and 90% empirical confidence interval ROC curves for Support Vector Classifier-based ensembles.

FY2008 Accomplishments and Results

The development of new ensemble classifier algorithms involves the optimization of performance metrics such as receiver operating characteristic (ROC) curves with respect to a variety of ensemble design factors. In FY2008, we completed a comprehensive study of these factors and their impact on classifier performance. Our development of classification algorithms leveraged a Hidden Signal Detection application in which false alarms are deemed extremely costly. These efforts ultimately led to the development of several groundbreaking ensemble classifiers, two peer-reviewed publications, and one provisional patent.

Built from many cost-sensitive Support Vector Classifiers (SVCs), our novel Cost-Sensitive Random Subspace Support Vector Classifier (CS-RS-SVC) ensemble significantly outperforms existing SVC ensembles built from non-cost-sensitive SVCs. It achieves a 55.3% detection rate on Hidden Signal Detection at 5.5x10^{-6} false alarm rate. This is a 15.5% relative improvement over an approach built using conventional SVCs (RS-SVC) and about three times better compared to a standard Bagged-SVC ensemble (Fig. 1).

We also significantly enhanced the state-of-the-art Random Forest (RF) classifier by developing variants in which node decisions are no longer constrained to be axis-aligned or linear, resulting in more fluid decision boundaries that better separate the classes (Fig. 2). This new classifier, called the Discriminant Random Forest (DRF), is 40% more

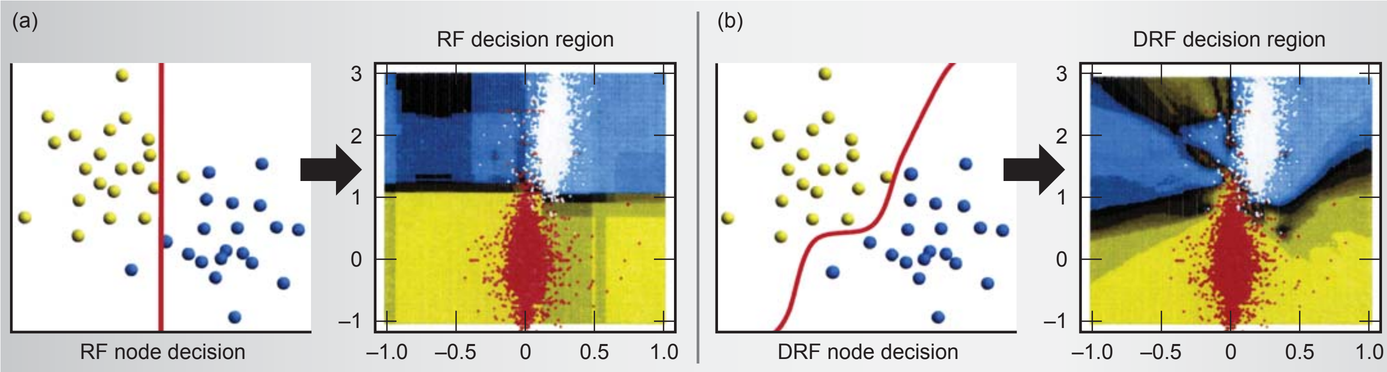

Figure 2. The Random Forest's (RF) axis-aligned linear decision boundaries in tree nodes give rise to "stair-step" decision regions (a), while the Discriminant Random Forest's (DRF) flexible node boundaries result in fluid, better-fitting decision regions (b).

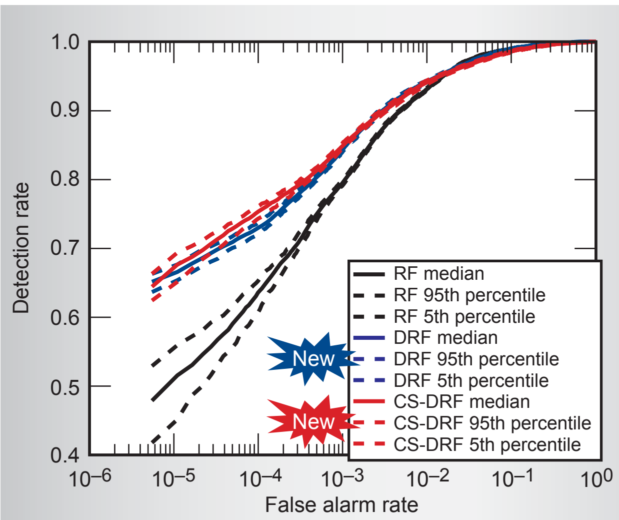

Figure 3. Median and 90% empirical confidence interval ROC curves for Random Forest-based ensembles.

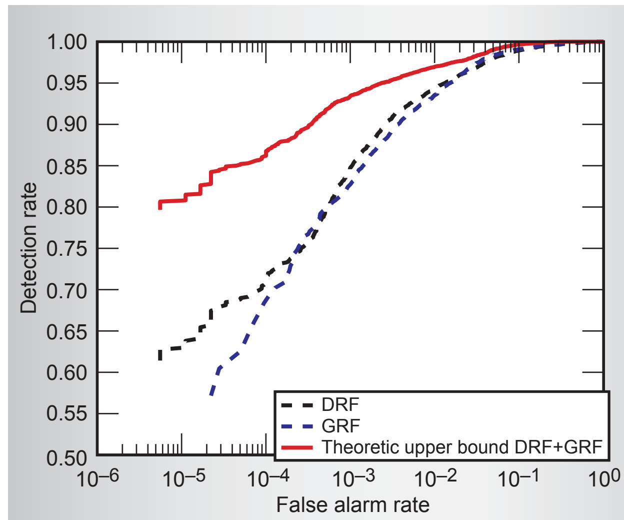

Figure 4. Two different Random Forest variants (DRF and GRF) are combined using an oracle, resulting in a theoretic upper bound on the performance of their combination.

compact, over six and a half times faster, and attains 36.2% higher detection rates at $5.5x10^{-6}$ false alarm rate than the conventional RF on the Hidden Signal Detection problem (65.1% versus 47.8%) (Fig 3). We also created a cost-sensitive extension to the DRF, the CS-DRF, which further improves DRF detection performance, particularly in the false alarm rate region around 10^{-4}.

Additionally, this research made substantial contributions to other LLNL detection applications, including Radiation Threat Detection (RadThreat) and Standoff High Explosives Detection (SHED). For RadThreat, our classifiers outperformed currently fielded approaches, even on heavily shielded sources. In SHED, our classifiers achieved significantly faster and higher detection rates at lower false alarm rates than human experts on the same data set.

Related References

1. Breiman, L., "Random Forests," *Machine Learning*, **45**, 1, pp. 5–32, 2001.
2. Davenport, M. A., R. G. Baraniuk, and C. D. Scott, "Controlling False Alarms with Support Vector Machines," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2006.
3. Ho, T. K., "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 8, pp. 832–844, 1998.
4. Lemmond, T. D., A. O. Hatch, B. Y. Chen, D. A. Knapp, L. J. Hiller, M. J. Mugge, and W. G. Hanley, "Discriminant Random Forests," *Proceedings of 2008 International Conference on Data Mining*, 2008.
5. Valentini, G., and T. G. Dietterich, "Bias-Variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods," *Journal of Machine Learning Research*, **5**, pp. 725–775, 2004.

FY2009 Proposed Work

In FY2009, we will 1) further enhance detection performance by extending our homogeneous ensembles to ensembles of heterogeneous base classifiers, exploiting the game-changing potential of combining different ensemble classifiers (Fig. 4); 2) extend the DRF methodology to allow more flexible node decisions; 3) generalize the current binary classifiers to handle multi-class situations; and 4) develop adaptive extensions of our learning approaches to address changing costs and changing data distributions.